Solutions

Exercise 1

Following is the Euclidean distance between A_i' and μ_1,

$$\begin{aligned}|A_i{'} - \mu_1|^2 &= (A_i' - \mu_1) \cdot (A_i' - \mu_1) \\ &= |A_i'|^2 + |\mu_1|^2 - 2A_i' \cdot \mu_1 \\ &= 2 - 2A_i' \cdot \mu_1\end{aligned}$$

A_i' is closer to μ_1 than to μ_2 means that

$$|A_i{'} - \mu_1|^2 < |A_i{'} - \mu_2|^2$$

use the first equation

$$\begin{aligned}2 - 2A_i' \cdot \mu_1 &< 2 - 2A_i' \cdot \mu_2 \\ 2A_i' \cdot \mu_1 &> 2A_i' \cdot \mu_2 \\ \frac{A_i}{|A_i|} \cdot \mu_1 &> \frac{A_i}{|A_i|} \cdot \mu_2 \\ \frac{D_{1i}}{|A_i|} &> \frac{D_{2i}}{|A_i|} \\ D_{1i} &> D_{2i}\end{aligned}$$

Q.E.D

Exercise 2

$$M_0 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Round 1

$$M_0^T \cdot A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Using the result from Exercise 1, we can conclude that doc 1, 2 belong to the first centroid, and doc 3, 4, 5 belong to the second, and it's easy to compute the new centroids

$$M_1 = \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{2}{3} \\ 0 & \frac{1}{3} \end{pmatrix}$$

Round 2

$$M_1^T \cdot A = \begin{pmatrix} 1 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} & 1 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{2}{3} & 1 \end{pmatrix}$$

Doc 1, 2, 3 belong to the first centroid, and doc 4, 5 belong to the second centroid. Compute the new centroids:

$$M_2 = \begin{pmatrix} \frac{2}{3} & 0 \\ \frac{2}{3} & 0 \\ 0 & 1 \\ 0 & \frac{1}{2} \end{pmatrix}$$

Round 3

$$M_2^T \cdot A = \begin{pmatrix} \frac{2}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{4}{3} & \frac{2}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{3}{2} \end{pmatrix}$$

Doc 1, 2, 3 belong to the first centroid and doc 4, 5 belong to the second, which is the same with round 2. The process have converged, and we're done. M_2 represents the final centroids.

Exercise 3

Start with μ_1 and μ_2 as the centroids of cluster 1 and 2, let's do step A:

$$\begin{pmatrix} \mu_1^T \\ \mu_2^T \end{pmatrix} \cdot \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} = \begin{pmatrix} \mu_1^T \cdot A_1 & 0 \\ 0 & \mu_2^T \cdot A_2 \end{pmatrix} = \begin{pmatrix} \mu_1^{'} & \mu_2^{'} \end{pmatrix}$$

It can be found that most documents in A_1 before will still belong to the first cluster, according to what we've got in exercise 1, so will the documents in A_2. It's possible that, due to the tie breaking strategy, some docs in A_1 can be clustered as members of cluster 2, but it won't have any change on the conclusion that $\mu_2^{'}$ is a linear combination of docs in the original cluster 2.

Because μ_1 (μ_2) and $\mu_1^{'}$ ($\mu_2^{'}$) both are linear combinations of docs in cluster 1 (2), we can conclude that the resulting clustering will not change anymore after the first step.